Parameter Estimation with Latent Variables

- Model \( p(\mathbf{X}, Z; \theta) \), observed data \( \mathbf{X} \), latent variables \( Z \), model parameters \( \theta \)
- Recall GMM, \( Z \): cluster assignments, \( \theta \): GMM parameters \( (\pi_k, \mu_k, \Sigma_k) \), \( k=1 \ldots K \)
- Goal: Estimate the model parameters \( \theta \) via MLE

\[
\hat{\theta} = \arg \max_{\theta} \log p(\mathbf{X}; \theta) = \arg \max_{\theta} \log \sum_Z p(\mathbf{X}, Z; \theta)
\]

- Doing MLE in such models can be difficult because of the log-sum

- If we "knew" \( Z \), sum over all possible \( Z \) not needed. Just define "complete data" \( (\mathbf{X}, Z) \), and do MLE on the complete data log-lik. \( \log p(\mathbf{X}, Z; \theta) \)

- Assumption: MLE on \( \log p(\mathbf{X}, Z; \theta) \) is easy
  - It often indeed is, especially when \( p(\mathbf{X}, Z; \theta) \) is exponential family distribution
    (or product of exponential family distributions)

Solution: An Iterative Scheme (EM Algorithm)

Initialize the parameters: \( \theta^{old} \). Then alternate between these steps:

- E (Expectation) step:
  - Compute the posterior \( p(Z|X, \theta^{old}) \) over latent variables \( Z \) using \( \theta^{old} \)
  - Compute the expected complete data log-likelihood w.r.t. this posterior

\[
Q(\theta; \theta^{old}) = \mathbb{E}_{Z \sim p(Z|X, \theta^{old})}[\log p(\mathbf{X}, Z; \theta)] = \sum_Z p(Z|X, \theta^{old}) \log p(Z, X; \theta)
\]

- M (Maximization) step:
  - Maximize the expected complete data log-likelihood w.r.t. \( \theta \)

\[
\theta^{new} = \arg \max_{\theta} Q(\theta; \theta^{old}) \quad (\text{if doing MLE})
\]

\[
\theta^{new} = \arg \max_{\theta} (Q(\theta; \theta^{old}) + \log p(\theta)) \quad (\text{if doing MAP})
\]

- If the log-likelihood or the parameter values not converged then set \( \theta^{old} \rightarrow \theta^{new} \) and go to the E step.

Why is this doing the right thing?
Illustration: EM for GMM

- Recall that the GMM parameters $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$
- The complete data likelihood
  
  $p(X, Z|\theta) = \prod_{n=1}^N \prod_{k=1}^K p(z_n = k)p(x_n|z_n = k) = \prod_{n=1}^N \prod_{k=1}^K \pi_k N(x_n|\mu_k, \Sigma_k)^{z_{nk}}$

- Taking the log, we get:
  
  $\log p(X, Z|\theta) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \log \pi_k + \log N(x_n|\mu_k, \Sigma_k)$

- E-step computes the expected complete data log-likelihood:
  
  $E_{Z|X, \theta}[\log p(X, Z|\theta)] = \sum_{n=1}^N \sum_{k=1}^K E[z_{nk}] [\log \pi_k + \log N(x_n|\mu_k, \Sigma_k)]$

  where $E[z_{nk}]$ is the expected value of $z_{nk}$ under the posterior

Illustration: EM for GMM (Contd.)

- The only expectation we need to compute $E_{Z|X, \theta}[\log p(X, Z|\theta)]$ is
  
  $E[z_{nk}] = \sum_{k'=1}^K \pi_{k'} p(x_n|k', \mu_k, \Sigma_k) = \pi_{k_0} p(x_n|k_0, \mu_k, \Sigma_k) = \frac{\pi_{k_0} N(x_n|\mu_{k_0}, \Sigma_k)}{\sum_{k=1}^K \pi_k N(x_n|\mu_k, \Sigma_k)} = \gamma_{nk}$

- Thus the expected complete data log-likelihood
  
  $E_{Z|X, \theta}[\log p(X, Z|\theta)] = \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} [\log \pi_k + \log N(x_n|\mu_k, \Sigma_k)]$

- M-step maximizes the the expected complete data log-likelihood w.r.t. $\pi_k, \mu_k, \Sigma_k$
- The update equations for these will be (shown on the board)
  
  $\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} x_n, \quad \Sigma_k = \frac{N_k - 1}{N_k - K} \sum_{n=1}^N \gamma_{nk} (X_n - \mu_k)(X_n - \mu_k)^	op, \quad \pi_k = \frac{N_k}{N}$

  where $N_k = \sum_{n=1}^N \gamma_{nk}$ is “effective” num. of examples assigned to $k$th Gaussian

Justification 1

- Consider the log likelihood on “incomplete” data $X$
  
  $\log p(X|\theta) = \log \sum_{Z} p(X, Z|\theta) = \log \sum_{Z} q(Z)p(X, Z|\theta)$
  
  \[ \geq \sum_{Z} q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} \] (using Jensen’s inequality for concave log)
  
  \[ \log p(X|\theta) \geq \sum_{Z} q(Z) \log p(X|Z, \theta) - \sum_{Z} q(Z) \log q(Z) = \sum_{Z} q(Z) \log p(X|Z, \theta) \] (const. doesn’t depend on $\theta$)

- If we set $q(Z) = p(Z|X, \theta)$, then the above inequality becomes equality

- Thus for $q(Z) = p(Z|X, \theta)$, we have
  
  $\log p(X|\theta) = \sum_{Z} q(Z) \log p(X|Z, \theta) \log p(X|\theta) + \log p(X|\theta)$

  Thus $\log p(X|\theta)$ is a tight lower-bound on $\log p(X|\theta)$

Why does EM work?
Justification 2

- We can also write the incomplete log likelihood

\[ \log p(X|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q||p_z) \]

where \( q \) is some distr. on \( Z \), \( p_z = p(Z|X, \theta) \) is the posterior over \( Z \), and

\[ \mathcal{L}(q, \theta) = \sum_x q(x) \log \left\{ \frac{p(x, Z|\theta)}{q(x)} \right\} \]

\[ \text{KL}(q||p_z) = -\sum_x q(x) \log \left\{ \frac{p(x, Z|\theta)}{q(x)} \right\} \]

(to verify, use \( \log p(x, Z|\theta) = \log p(x|Z, \theta) + \log p(Z|\theta) \) in the expression of \( \mathcal{L}(q, \theta) \))

- Since \( \text{KL}(q||p_z) \geq 0 \), \( \mathcal{L}(q, \theta) \) is a lower-bound on \( \log p(X|\theta) \) for any \( q \)

Justification 2 (contd.)

Recall \( \log p(X|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q||p_z) \). EM can also be seen as:

- With \( \theta \) fixed to \( \theta^{old} \), maximize \( \mathcal{L}(q, \theta^{old}) \) w.r.t. \( q \)

\[ \hat{q} = \arg \max_q \mathcal{L}(q, \theta^{old}) \]

which is equivalent to making \( \text{KL}(q||p_z) = 0 \) or setting \( \hat{q} = p(Z|X, \theta^{old}) \)

(This step makes \( \mathcal{L}(\hat{q}, \theta^{old}) = \log p(X|\theta^{old}) \); see next slide)

- With \( \hat{q} \) fixed at \( p(Z|X, \theta^{old}) \), maximize \( \mathcal{L}(\hat{q}, \theta) \) w.r.t. \( \theta \), where

\[ \mathcal{L}(\hat{q}, \theta) = \sum_x p(x, \theta^{old}) \log p(x, Z|\theta) - \sum_x p(x, \theta^{old}) \log p(x|\theta^{old}) \]

\[ = \frac{\log p(X|\theta^{old})}{\text{const.}} + \text{const} \]

\[ \theta^{new} = \arg \max_{\theta} \mathcal{L}(\hat{q}, \theta) \]

(This step ensures that \( \log p(X|\theta^{new}) \geq \log p(X|\theta^{old}) \); see next slide)

A View in the Parameter Space

- E-step: Update of \( q \) makes the \( \mathcal{L}(q, \theta) \) curve touch the \( \log p(X|\theta) \) curve
- M-step gives the maximum \( \theta^{new} \) of \( \mathcal{L}(q, \theta) \)
- Next E-step readsjusts \( \mathcal{L}(q, \theta) \) curve (green) to meet \( \log p(X|\theta) \) curve again
- This continues until a local maxima of \( \log p(X|\theta) \) is reached

Thus the E and M steps never decrease the log-likelihood \( p(X|\theta) \)
Some EM Variants

- **Generalized EM**: M step doesn’t require maximization w.r.t. \( \theta \); even if the new \( \theta \) just increases the lower bound, we will still converge to a local optima.

- **Variational EM and MCMC EM**: If the E step of computing the posterior \( p(Z|X, \theta) \) is intractable, we can use variational Bayes (VB) or MCMC to approximate the posterior.

- **Expectation Conditional Maximization**: Parameters are partitioned in groups. M step consists of multiple steps (each optimizing one group of parameters, treating all other groups as fixed).

- **Online/incremental EM**: E step only processes one (or a small number of) observation, computing posteriors/expectations only w.r.t. that minibatch of data. For exponential family distributions, the sufficient statistics needed in the M step can be easily updated incrementally, leading to simple form of incremental parameter updates. Very useful for scalable inference. See:
  1. Online EM Algorithm for Latent Data Models (Cappé & Moulines, 2009)
  2. Online EM for Unsupervised Models (Liang & Klein, 2009)

Next up: Probabilistic PCA and Factor Analysis